International Journal Of Artificial Intelegence Research
Vol 2, No 1, June 2018, pp. 18-27
ISSN: 2579-7298

The Comparison of Signature Verification Result Using 2DPCA Method and SSE Method

Anita Sindar RM Sinaga

Teknik Informatika, STMIK Penusa Medan, Indonesia
haito_ita@yahoo.com

ARTICLE INFO

Article history:
Received: 2017-12-21
Revised: 2018-01-18
Accepted: 2018-02-06

Keywords:
Signature
2DPCA Method
SSE Method
Comparison
Accuracy

ABSTRACT

Everyone has signature characteristics, but it will be challenging to match original signatures with a clone. Two dimensional Principal Component Analysis (2DPCA) method, Sum Equal Error (SSE) method includes a plan that can provide accurate data verification value of 90% - 98%. The initial process of both ways has the equation that is using the digital image, the results of scanned signatures resize the conversion to grayscale. The objective of comparing the two approaches is to find a more accurate and precise way of identifying a name and determining the appropriate signature pattern for both methods. The extraction process of each method requires experimental data as a data source in pixel size. A digital image consists of a collection of pixels then each image is converted in a matrix. Preprocessing Method 2 DPCA each data is divided into data planning and data testing. Extraction on SSE method, each information sought histogram value and total black value. This study yields a comparison of the suitability of the extraction results of each process. Both of these methods have a data accuracy rate of 97%-98%. When compared to the results of the accuracy of image verification with 2DPCA method: SSE is 97 %: 96%. The difference in the accuracy of identifying signatures between 2DPCA method with SSE method is 2%-3% influenced by thickness, sharpness, and complexity of a stamp.

I. Introduction

The identification mark is used to indicate a person's identity. Each opening of the account number of the consumer is required to put the signature according to the ID card. It is highly probable that trademarks can be easily imitated without arousing suspicion. Technological developments facilitate data manipulation with a legality rate of about 80%-90%. Manually manipulating the signature is easy by watching the hand gestures directly, following the ink streak and can also use the scanner and then imitated. Trademarks are handwritten forms that contain special characters and additional structures that are often used as proof of identity verification. Manual signature pattern recognition is done with repeated stamps, this process takes a long time, and the results are not satisfactory. Matching signature characteristics using a computer, aims to produce an accuracy of information that is closer to the authenticity of the image (signature).

Principal Component Analysis (PCA) method to reduce the dimension of input data variable into a smaller dimensionless main component with minimum information loss, the central part formed is not correlated to one another while Sum Squared Error (SSE) method, finds the quadratic value of error difference from sample data and test data. Two-dimensional Principal Component Analysis (2DP-CA) has been widely used for image representation and recognition [1].

The scope of image processing and statistical pattern recognition model is a deductive basis for building a conceptual framework that explains the process of empirical data signatures. It takes some original names, scan results and image processing techniques to be able to identify matching titles. Patterns and features of great name, complexity, and pressure.

The initial process of identifying the signatures of the 2DPCA method and the SSE method using a digital sample image which is starting converted into a grayscale image. Both ways produce an identification level with high accuracy and an uncomplicated verification phase. Both of these methods are perfect for a thick and uncomplicated signature style. The 2DPCA method is a feature extraction for data compression and requires many coefficients in storing data [2].

For the color process using the first sum squared error (SSE) method is to make the template database as the comparison with the value of the crop image processing, the result of the comparison, from the comparison result will be obtained SSE blue, SSE green value in each sample data. By searching the squared value of error difference from sample data and test data. The results of the test show that SSE can recognize signatures with 96% accuracy. 2DPCA Method used for signature extraction with various slopes as well using Euclidean Distance to look for signature similarities [3]. This study aims to find out which method is more appropriate in verifying and knowing the proper signature pattern for 2DPCA Method and SSE Method.

II. Related Work

Outline the comparison stage of the signature matching extraction results the 2DPCA method and SSE method as follow a. Preprocessing, b. Extract 2DPCA method c. Euclidean Distance techniques d. Extract SSE method. Digital image extraction begins by converting the image into rows and matrices columns. The model contains colors Red, Blue, and Green. Image representation results in the matrix will facilitate image reduction. The data source consists of several signatures affixed to the white paper to be sampled data. The results of the signature capture are stored in the.JPG file format. The pattern is a grouping of numerical and symbolic data (including imagery) automatically by a machine (computer). Operation pattern recognition system:

a) Stage of exercise

It consists of feature extraction design, decision rule design, pattern recognition evaluation, and knowledge data formation.

b) Introduction phase (operational)

c) Consists of determining the pattern to be observed, the measurement of features, the process of recognition by enacting the rules of decision and the use of known data.

An image must be represented numerically with discrete values. The image representation of the malar function (continuous) into discrete values is called digitalization. The resulting image is called digital image (digital image). In general, Digital images are four rectangles, and their dimensions are expressed as height x width (or width x length). Digital images of the size of m x n are commonly represented by the matrix of the row size m and n columns as follows. On average, the feature matrices produced by the proposed method and those formed by 2PCA are about the same size. A digital image is regarded as the numeric representation of the 2D image in a sampled and quantized from. The basic picture element is called pixel, and an MXN image has M rows of pixels and N columns of pixels. With a coordinate system that follows the principle of displacement on a standard TV screen, a pixel has coordinates of (x, y), x represents column position and y denotes the line position. The upper-left corner pixel has coordinates (0, 0) and the pixels in the lower-right corner have coordinates (N-1, M-1). It can also be thought of a 2D grid or matrix whose elements are represented by f(x,y), where x and y are the coordinates of the grid or the indices of the matrix element [4]:

$$f(x) = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,N) \\ f(0,1) & f(1,1) & \dots & f(1,N) \\ \vdots & \vdots & \vdots & \vdots \\ f(M-1,0) & f(N-1,1) & \dots & f(N-1,M-1) \end{bmatrix}$$

Principal Components Analysis (PCA) or Karhunen Loeve Transformation is a technique used to simplify a data, using a linear transformation to form a new coordinate system with maximum variance. Two Dimensional Principal Component Analysis (2DPCA) is a development of the

Principal Component Analysis (PCA) method that serves feature extraction for data compression. The 2DPCA method has advantages of the PCA method regarding data accuracy and time complexity but has the disadvantage of requiring multiple coefficients in storing data [5]. Many 2DPCA-based face recognition approaches pay a lot of attention to the feature extraction but fail to pay necessary attention to the classification measures. The typical classification measure used in 2DPCA-based face recognition is the sum of the Euclidean distance between two feature vectors in a feature matrix, called distance measure (DM).

The 2DPCA method is defined as follows: The 2DPCA method if an image matrix (m x n) then the pattern of the image does not need to be transformed into a one-dimensional model. An image A with a design of sizes (m x n) and X denotes an n-dimensional unity column vector. To project the image A, (M x N) to X matrix with a linear transformation.

$$2D - PCA = AX \quad \text{.....(1)}$$

After getting the matrix of image X then the next step normalization matrix. Then calculate the mean matrix (μ) to obtain the center matrix. To get the zero mean (Φ) the value of μ which is the mean matrix value is given in equation (2) [6].

$$\Phi j, i = xj, i - \mu \quad \text{.....(2)}$$

From the calculation result zero mean is used to get the value of covariance matrix *(C)* by switching the zero mean transpose. To obtain the characteristics of a sample represented in matrix form, it is calculated and eigenvalue of the covariance matrix. If C is a square matrix of any size m > 1, then the non zero vector Λ in R_n is called the eigenvector of C if CΛ scalar multiplier of Λ, calculated using equation (3) [7].

$$C\Lambda = \lambda\Lambda \quad \text{.....(3)}$$

C: covariance matrix

Λ: eigenvector

λ: eigenvalue

Scalar λ referred to as Scalar λ is called the eigenvalue of C and Λ is called the eigenvector of C corresponding to λ. To obtain the eigenvector (Λ) and eigenvalue (λ), then from equation (3) can be written into (4).

$$C \Lambda = \lambda I \Lambda$$

$$(\lambda I - C) \Lambda = 0$$

$$\text{Det}(\lambda - C)\Lambda = 0 \quad \text{.....(4)}$$

The result of equation (4) is a matrix, i.e., eigenvalue (λ) sequentially decreased from the most considerable value to the smallest value (λ1> λ2> λ3 ..... λm). The eigenvector (Λ) corresponding to the most significant amount of the eigenvalue has the most dominant character, while the eigenvector value corresponding to the smallest eigenvalue has the least dominant feature.

2.3. Euclidean Distance

Euclidean Distance is one of the image matching techniques that is by the method of adjacent neighbor classification by calculating the roots of squares of the difference between 2 vectors. Euclidean distance matrices (EDMs) are matrices of the squared distances between points [8]. Euclidean Distance formula is written as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(X_{ik} - X_{ij})^2} \text{.....(5)}$$

d_{ij} : *Euclidean* distance between *i* and *j*

x_{ik}: data training

x_{ij} : data testing

n: amount data training

2.4. Sum Square Error (SSE) Method

Sum Square Error (SSE) Method is one of the statistical methods used to measure the total difference of the actual value against the achieved value. SSE is also called Summed Square of Residuals. SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster [9].

$$SSE = \sum_{i-1}^{n}(x_i - y_i)^2 \text{.....(6)}$$

x = actual or actual value y = achieved value

The value of X in this study is the value stored in the database while the value of Y is a component of test data. The SSE method value near 0 indicates that the model has the smallest random error component and that value will be more useful for forecasting against an observed model.

III. Prepare Your Paper Before Styling Results and Discussions

Every signature is having unique characters like bold, thin scratches and complicated. Data sources are signature collection from 10 sample writes in the white paper. Before the data is used, pre-process stage to facilitate the next step. At this stage, the process is to resize the image of the signature data from the original size (Figure 1).

106 x 244 | 108 x 142 | 90 x 27

Image 1 | Image 2 | Image 3

Fig. 1. Results Scanned of Original Signature Size (Pixel)

3.1. Preprocessing

For input data testing and data training, the first, each data is done resizing of image to facilitate the identification of model into matrix form. In the sample, an image resizes to 100 x 100 pixel (Figure 2).

Fig. 2. Resize Image 100 x 100 Pixel

The RGB image converted to Grayscale is stored in an 8-bit format for each sample pixels allowing as many as 256 intensities. A grayscale image is a digital image that has only one channel value on each pixel, meaning the value of Red and Blue. The process of image processing training and testing has the same path. Input Data in the form of signature images processed by the grayscaling process. An

image intensity value greater than or equal to the threshold value will be changed to 1 (white) while the image intensity value less than the threshold value will be changed to 0 (black). So that the output image of thresholding result is in the form of a binary image. The equations used to convert the grayscale image pixels value to binary in the thresholding method are:

$$g(x, y) = \begin{cases} 1, \textit{ jika } f(x,y) \geq \mathrm{T} \\ 0, \textit{ jika } f(x,y) < \mathrm{T} \end{cases}$$

f(x,y) : grayscale image
g(x,y) : binary image
T: threshold value

Binary imagery is called Black and White image or monochrome image. It takes 1 bit to represent each pixel of the binary image. Image processing is divided into two parts that are testing or insert image data to be used as reference process identification and process second is training or testing system.

Fig. 3.Thresholding Results

3.2. Implementation of 2DPCA Method

The signature data is collected and begins with the RGB image conversion process - grayscale - thresholding. The image of the thresholding process is represented in the form of a matrix of sizes (M x N), and X denotes an n-dimensional unity column vector. Grayscale images contain fewer values of 8-bit colors than RGB pictures with 24-bit colors. Standard monochrome (black-and-white) imagery, with variations in intensity from 0 to 255, in color images there are 16,777,216 color variations when each component R, G, and B contains 256 levels of energy. Data training is represented in the matrix (n x m) as much as the amount of data. The image of thresholding result is expressed in ASCI number (Figure3).

255	255	255	255	255	255	255	255	255	255	255	255	255	255	255
250	250	255	255	249	254	255	255	252	9	2	0	12	255	254
255	255	243	255	255	252	248	0	0	0	7	3	0	255	250
239	249	253	246	251	255	0	0	0	15	0	0	0	255	251	255
255	255	255	255	247	3	0	8	0	255	8	11	0	249	255	246
244	253	255	243	4	0	0	2	255	247	0	3	0	255	255	255
255	255	245	11	17	1	0	254	254	255	2	0	4	248	255	249
250	244	15	0	0	9	255	255	245	247	2	0	0	255	238	255
255	0	0	5	255	236	255	253	255	255	1	0	255	255	251	255
0	10	0	9	251	255	253	252	255	0	2	3	248	253	248	255
4	0	255	250	255	246	255	255	242	23	0	2	244	255	255	245

Fig. 4.Image Representation (:,:, 1) in Matrix

The RGB image is a type of image that consists of colors in the form of components R (red), G (green), and B (blue). Each color component uses 8 bits (its value ranges from 0 to 255). Possible colors that can be produced reach 255x255x255 or 16,581,375 colors.

Table 1. Data Source in Matrix

255	249	246	3	5	255	255	255	255	247
246	255	251	0	8	247	248	255	247	252
2	0	0	255	248	255	252	252	243	255
12	0	3	248	255	246	255	255	252	244
0	11	0	15	1	255	253	245	255	255
13	3	3	0	0	0	255	255	255	244
0	0	2	7	21	0	1	255	255	246
1	4	0	0	4	0	4	250	244	255
13	0	4	12	0	24	3	1	0	255
0	6	0	0	7	0	0	0	2	252

Table 1 is a piece of data source testing image 2 in the form of matrix representation. The extraction process of 2DPCA data testing begins with the search of X (column vector) of the matrix (n x m). Next search Mean (μ) from matrix. ($\mu = x_{1,i} + x_{2,i} + x_{3,i} + \ldots + x_{m,i}/ m$) dan Zero Mean (Φ). Result Covariance Matrix *(C)* diperoleh dari $C = 1 / m - 1 (x_{ji} - \mu_i) (x_{ji} - \mu_i)$ (Table 2).

Table 2. Result Covariance Matrix

2.579479	0.436079	0.116479	0.002279	-0.052621	-0.083221	-0.102021
2.573079	0.432379	0.113379	-0.000421	-0.055021	-0.085521	-0.116521
2.442779	0.372479	0.060279	-0.053721	-0.110221	-0.141221	-0.163321
2.455379	0.376679	0.063079	-0.051621	-0.108421	-0.141221	-0.162021
2.193579	0.270679	-0.022721	-0.132521	-0.188821	-0.141221	-0.245421
2.200679	0.272979	-0.021121	-0.131221	-0.187621	-0.222821	-0.244221
1.430079	-0.007221	-0.227321	-0.310321	-0.353321	-0.221521	-0.409421
1.391179	-0.020721	-0.236921	-0.318521	-0.360721	-0.379421	-0.409421
0.942379	-0.155021	-0.318821	-0.378021	-0.407321	-0.386321	-0.443421
0.414479	-0.355721	-0.473821	-0.518421	-0.541621	-0.555621	-0.571821

Data testing is stored in the database to be further used as a source of data comparison with training data. Eigenvalue (λ) is sequenced in descending order from the most considerable value to the smallest value (λ1> λ2> λ3 ..... λm).

The Eigenvector (Λ) corresponding to the most significant value of the eigenvalue has the most dominant character, while the eigenvector value corresponding to the smallest eigenvalue has the least dominant feature. The optimal eigenvector is called the projection matrix (system). The projection matrix will be used for the formation of the final pattern of data, i.e., as the method of recognition method for the design of data testing. Table 2 is the Eigenvalue (λ) and Eigenvector (Λ) of the data testing piece.

Table 3. Eigenvector and Eigenvalue of Results

Eigenvalue	Eigenvector
-0.0375 + 0.0227i	6.8888 + 0.0015i
-0.0375 - 0.0227i	6.8780 + 0.0016i
-0.0020 + 0.0128i	6.6215 + 0.0016i
-0.0020 - 0.0128i	6.6502 + 0.0016i
3.0753	6.0997 + 0.0026i
0.1269	6.1162 + 0.0016i
0	4.4179 + 0.0034i
-0.0003	4.3310 + 0.0013i
-0.012	3.2719 + 0.0014i
-2.4361	2.1301 + 0.0007i

From the eigenvector and eigenvalue values a new set of data sets (NewDataSet = ΦT * Λ) and the matrix score (MatWeight = X * NewDataSetT) are matched. This 2DPCA method extraction step results in training data in the matrix. The final step is to perform a signature match that is to measure the matrix score and the projection matrix of data testing that has been projected with the projection data training matrix. Matching signatures using Euclidean Distance (Table 2), ie, finding the Euclidean distance between i and j. The smallest Euclidean Distance (dij) weight becomes the source of the image identifying data.

Table 4. Euclidean Distance of Results

Data 1	Data 2	Data 3	Data 4	Data 5
23519.60	48705.63	54651.06	46126.79	52781.17

Data verification uses five scenarios of 100 data, each situation consisting of 50 training data and 50 data testing and ten signature data classes. The percentage comparison of training and testing data is 50%: 50%.

Table 5. Verification Test Results

Scenario	**Data Size**	**Accuracy**	**Time (s)**
1	50 x 50	92%	0.5782
2	100 x 100	94%	0.1654
3	150 x 150	97%	0.2756
4	200 x 200	98%	0.3645
5	250 x 250	95%	0.4749

Graph of matching signature based on data testing and training data on the smallest weight of Euclidean Distance (dij), according to a diagram obtained data testing and training data follow Euclidean Distance Weight.

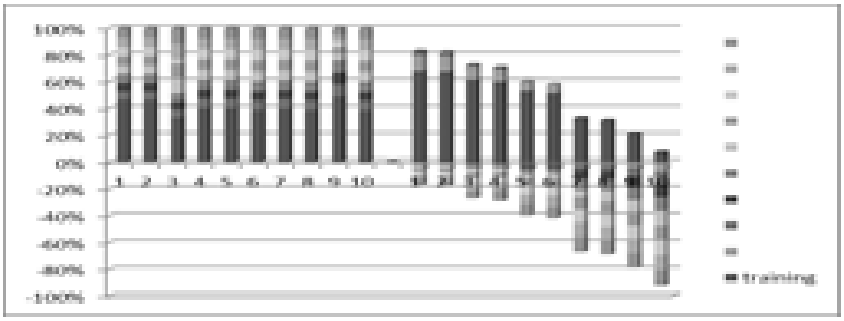

Fig. 5. Data Graph Testing and Data Training

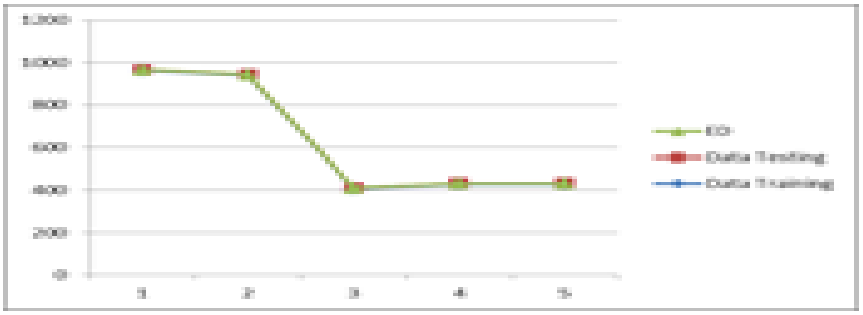

Fig. 6. Signature Matching Graph

An image histogram is a graph depicting the deployment of pixel intensity values of a particular image or part in an image. From a histogram can be known the frequency of emergence relative (relative) of the intensity of the image.

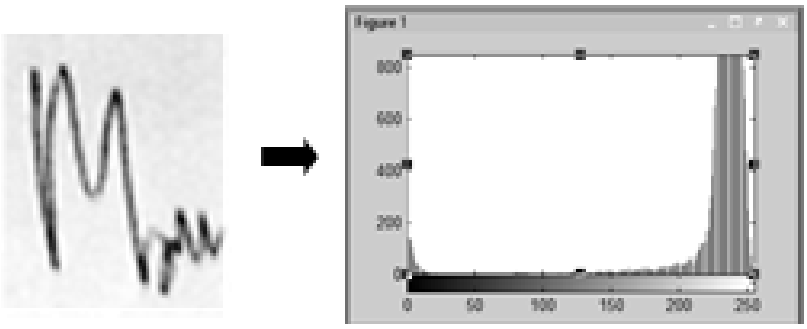

Fig. 7. Image Histogram 240 x 160 Pixels

The histogram also shows the brightness and contrast of an image. Suppose that the digital image has an L of gray, that is, from 0 to L - 1 (for example in an image with quantization of 8-bit grays degree, a gray degree value from 0 to 255). Mathematically hn = ni / n, i = 0.1, ... L - 1; ni = the number of pixels that have the gray degree L, n = the total number of pixels in the image. Sum Squared Error (SSE) to determine the value of the histogram and the total black values of the sample data of the scanned test data

Table 6. Result of Histogram Value and Total Black Value

Data	Value Histogram	Total Black Value
1	0,05783	95%
2	1,073842	98%
3	0,329321	96%
4	0,923247	97%
5	1,839358	94%

From the result of histogram value and total black value (%) determined SSE value that is a difference of actual value with value achieved. SSE results show the accuracy of the signature.

Table 7. Value Result of SSE Method

Data	Size Data	Accuracy	Time (s)
1	50 x 50	95%	0.267836
2	100 x 100	97%	0.199104
3	150 x 150	93%	0.333916
4	200 x 200	96%	0.574312
5	250 x 250	91%	0.464381

Implementation of each result The stored method then makes a comparison of the accuracy of the 2DPCA method results with the SSE method.

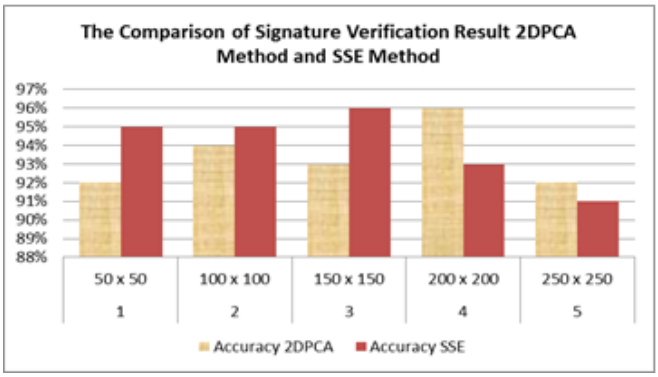

Fig. 8. The Comparison Diagram of the 2DPCA

The sample data is processed by applying the 2DPCA and SSE methods and continued matching of the signature. The 2DPCA reduction method is a feature extraction process that can reduce the dimensions of the image to be smaller. The results of the implementation and testing stages will be analyzed to determine the accuracy of the method used. The DPCA 2 method is more appropriate to match signatures than the SSE method, influenced by the thickness, complexity of a name and the quality of the captured camera image or the resulting digital image scan.

IV. Conclusion

The 2DPCA method represents the image with a much smaller covariant Matrix, so the evaluation of the covariance matrix is more accurate. Time to calculate eigenvalues and eigenvectors faster. The use of different data training and testing aims to determine the effect of using the amount of training data on accuracy. The result of Euclidean Distance calculation, the smaller the distance between the score matrix with the projection matrix of data testing that has been projected with the projection data training matrix, the closer the distance to the matching image. By using Euclidean Distance calculation, identification of sample data with 2DPCA method produces accuracy of 96%-98%. The data identification process using the Sum Squared Error (SSE) range value has an efficiency of 95% - 98%. Supported data sample has dominant color or thickness. The comparison results of both methods have a reasonably good accuracy level in the signature matching of about 97% - 98%. In this research, signature matching is done using 2DPCA reduction method and SSE method. To identify the signature match with the Sum Squared Error (SSE) Method, Sample Data, SSE Value Data and Test Data, as well as 2DPCA method required data testing, data planning, and Euclidean Distance Method. The 2DPCA method more accurately identifies the signature than the SSE method, the matching difference being 2% -3%.

Acknowledgment

I thank STMIK Penusa Medan for entrusting me to teach Image Processing lesson. I think it is a chance to occupy the interest sector of my capability. Hopefully, I can transfer my skill both for teaching and do research entirely.

References

[1] Qian Wang, Qian Gao, "IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops" pp. 1152-1158, 2016.

[2] A. Mashhoori, M. Zolghadri Jahromi, "Block-wise two-directional 2DPCA with ensemble learning for face recognition", pp. 111-117, 2013.

[3] John Huguenard, "Stanford University School of Medicine," 2015

[4] E. Jebamalar Leavline, D. Asir Antony Gnana Singh, "On Teaching Digital Image Processing with MATLAB", American Journal of Signaling Processing, pp. 7-15, 2014.

[5] Xiuhua Liang, Guangming Deng, Bin Yan, "Fruit and Vegetable Nutrition Value Assessment and Replacement Based on the Principal Component Analysis and Cluster Analysis," Scientific Research Publishing, Applied Mathematics, pp. 1620-1629, 2016.

[6] Shoichi Fujimori, Yu Kawakami, Masatoshi Kokubu, Wayne Rossman, Masaaki Umehara, Kotaro Yamada, "Entire Zero-Mean Curvature Graphs I Lorents-Minkowski 3-Space, The Quarterly Journal of Mathematics, pp. 801-837, 2016.

[7] Santiago Velasco-Forero, Marcus Chen, Alvina Goh, "Comparative Analysis of Covariance Matrix Estimation for Anomaly Detection in Hyperspectral Image, IEEE Journal of Selected Topics in Signal Processing, pp.1061-1073, 2015.

[8] Ivan Dokmanic, "Euclidean Distance Matrices: Essential Theory, Algorithms, and Applications", IEEE Signal Processing Magazine, pp. 12-30, 2015.

[9] Tippaya Thinsungnoen, Nuntawut Kaoungku, PangsakomDurongdumronchai, "The Clustering Validity with Silhouette and Sum of Squared Errors', Proceedings of the 3rd International Conference on Industrial Application Engineering, pp.44-51, 2015